Supplementary Materials

1.1. Changes to the It's About Time questionnaire

A number of changes were made to adapt the questionnaire for adult self-report. Five items were removed from the questionnaire: How often does your child ask question about their past? How often does your child ask you questions about things that will happen in the future? How often does your child have difficulty waiting a while before he or she is allowed to do something they really would like to do now? On school morning, how often is your child ready to leave on time? In general, compared to other children of their age, how well developed is your child's sense of time?

Each item was reworded for self-report (e.g. rewording At bedtime, 'how often does your child talk about or seem to think about what has happened to them that day?' to 'when lying in bed, how often do you think about what has happened to you that day?'

The scale was changed from Rarely (0), Sometimes (1), Most of the Time (2) and Almost Always (3) to Never (0), Rarely (1), Sometimes (2), Often (3), Always (4).

1.2. Participant removal

Participant data was removed according to criteria specified in the pre-registration document and is detailed below for each task.

1.2.1. Temporal Difference Threshold.

The final sample size in the visual condition was Autistic n = 58, NT n = 90 and in the auditory condition Autistic n = 58, NT n = 91. Participants threshold were taken as the lowest estimate from the two runs of the staircase. Where the staircase had failed to settle between the 30th and 50th trial (the difference between the steps was +200ms or – 600ms) the run producing the lowest estimate of threshold was rejected. In the visual condition, 24 participants lowest thresholds were rejected according to this criterion. For 19 of these participants, the threshold on the run producing the higher estimate passed this criterion. Data was thus removed for one autistic and four NT participants where neither run met this inclusion criterion. In the auditory condition, 26 participants' lowest thresholds were rejected. For 24 of these participants, the threshold on the run producing the higher estimate passed this criterion. Data was removed for two NT participants where neither run met this inclusion criterion.

1.2.2. Verbal Estimation.

The final sample size in the visual condition was Autistic n = 56, NT n = 86 and in the auditory condition Autistic n = 55, NT n = 89. There were two inclusion criteria. Firstly, the fit of the linear regression needed to have a positive slope. This led to the removal of two NT participants in the visual condition. Second, to confirm the slope of the regression was significantly greater that 0 (α = .10). This led to the removal of three autistic and three NT participants in the visual condition and one autistic and three NT participants in the auditory condition.

1.2.3. Temporal Generalisation.

The final sample size in the Auditory 400ms Standard condition was Autistic n = 53, NT n = 87, in the Auditory 800 ms standard condition Autistic n = 56, NT = 89, in the Visual 400 ms standard condition Autistic n = 54, NT = 80, in the Visual 800ms standard condition Autistic n = 55, NT = 89, in the Pitch 500 Hz standard condition Autistic n = 54, NT n = 84 and in the Pitch 1000 Hz standard condition Autistic n = 49, NT n = 83. To confirm that participants could perform the task appropriately, responding *yes* to the stimuli closest to the duration of the standard, individual participant data in each condition was fitted using a quadratic regression to confirm that their temporal generalisation gradient approximated a parabola. Participants data was included where the regression model was significant ($\alpha < .04$[1]).

2.1. Analysis

2.1.1. Temporal Discrimination

Figure S1 displays mean trial level data across the 50 trials (the best run which was used to calculate the participant's threshold) on the temporal discrimination task

[1] Note that the pre-registered criteria was α = .010 similar to linear regression for VE task. However, many more data points are required for quadratic regression and using α =.010 as criteria led to the removal of a large number of participants who produced temporal generalisation gradients approximating a quadratic form.

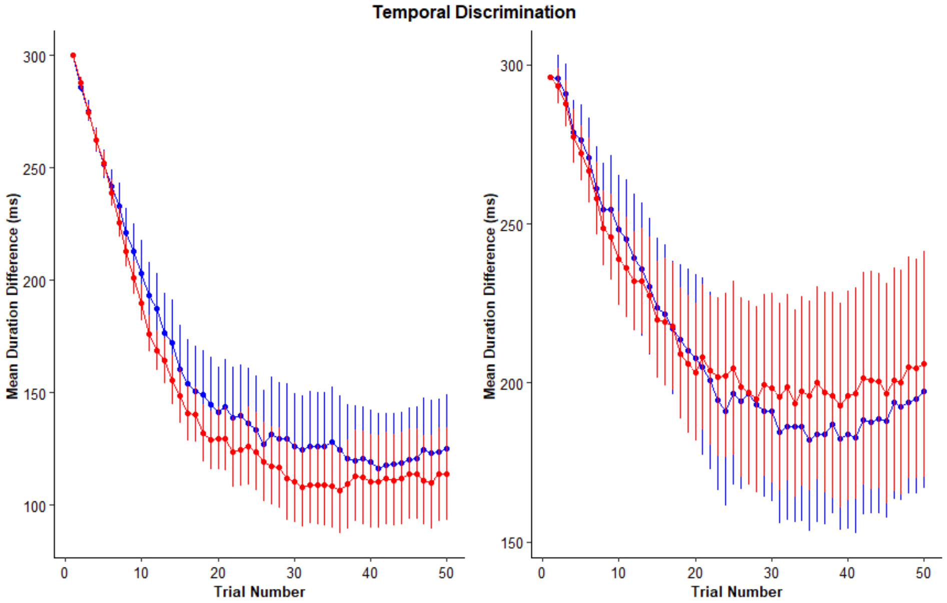

Figure S1 Performance across the best run of the staircase procedure for autistic (red) and non-autistic participants (blue). Data points give the mean difference between the durations on each trial. Error bars denote 95% CI.

2.1.2. Verbal Estimation Task

Participants mean estimates and coefficient of variation (mean – standard deviation) in response to each duration are represented in Figure S2

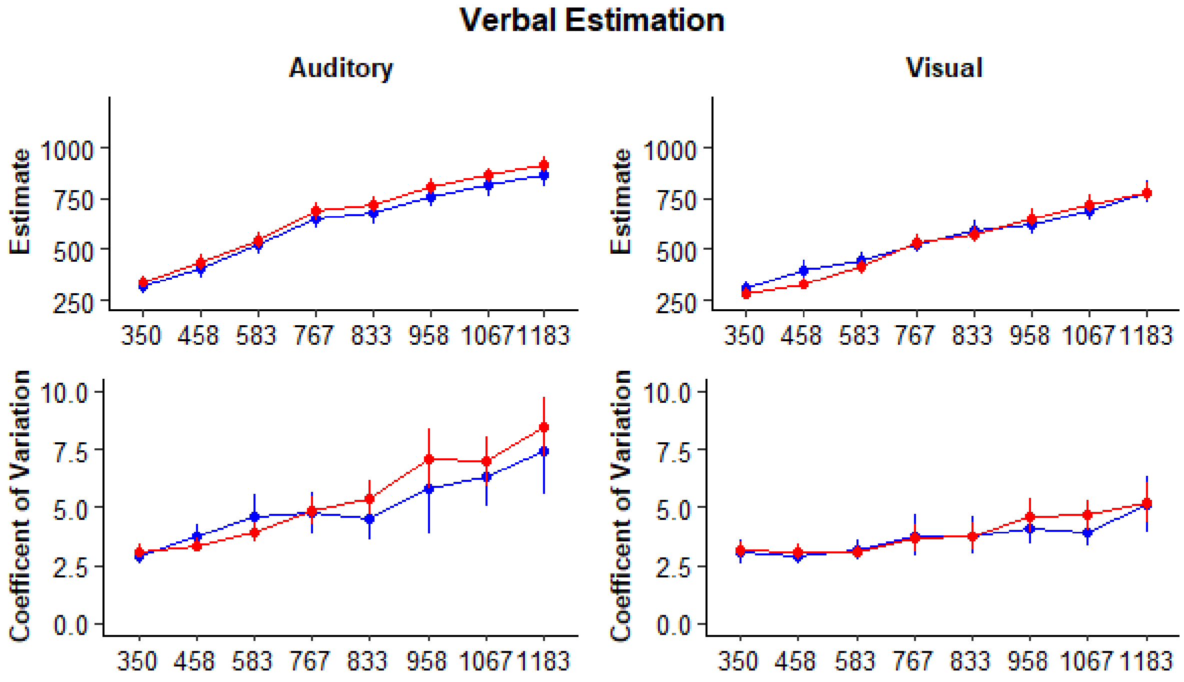

Figure S2 Mean estimates and coefficient of variation (mean – standard deviation) for each duration for the Autistic (red) and NT (blue) groups. Error bars denote 95% CI.

2.1.3. Temporal Generalisation Task

2.1.3.1. Bias

Bppd data for each condition is displayed in Figure S3.

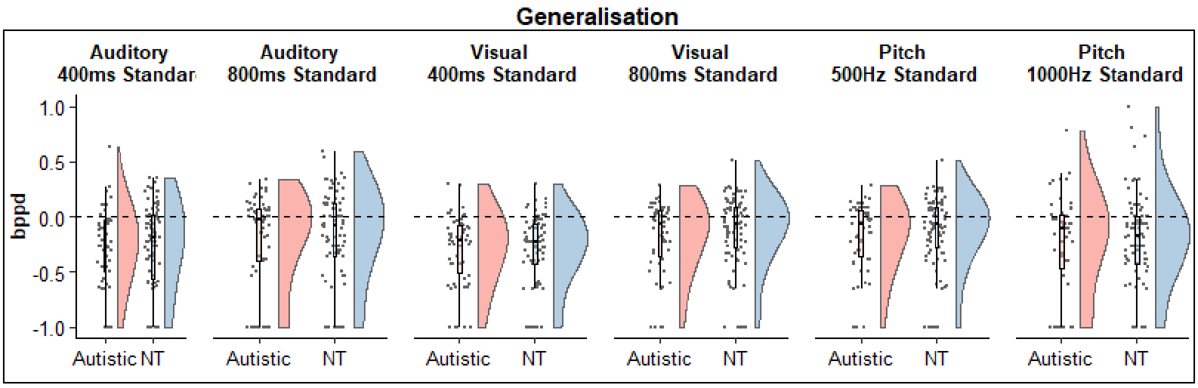

Figure S3. Raincloud plots displaying bppd (bias) on the Generalisation tasks (Temporal and Pitch) for autistic (pink) and neurotypical (NT; blue) participants. More negative values indicate participants were more likely to respond 'Yes' and more positive 'No'.

Temporal Generalisation: There was no main effect of Modality (F (1,123) = 0.35, *p* = .555, η^2= .003), nor Group (F (1, 123) = 0.33, *p* = .568, η^2= .003). There was a main effect of Standard (F (1, 123) = 31.28, *p* < .001, η^2= .203) indicating that bppd was more negative when responding to 400ms standards (mean = -0.29, SD = 0.35) compared to 800ms (mean = -0.17, SD = 0.39). More negative values indicate that participants were more likely to respond 'yes' to comparison stimuli. None of the other main effects or interactions reached statistical significance (F < 1.24, *p* > .268). The Bayes Factor for the between group comparisons of bppd in the visual 400 ms standard condition (BF_{10} = 0.20), for the visual 800 ms standard condition (BF_{10} = 0.27), the auditory 400 ms standard (BF_{10} = 0.19) and the auditory 800 ms standard (BF_{10} = 0.23) suggested that the data supported the null hypothesis.

Pitch Generalisation: There was no main effect of Standard (F (1,121) = 0.52, *p* = .474, η^2= .004) or Group (F (1,121) = 1.69, *p* = .196, η^2 = .014), or interaction (F (1, 121) = 0.19, *p* = .660, η^2= .002). The Bayes Factor for the between group comparisons of bppd for the 500 Hz standard (BF_{10} = 0.39) and for the 1000 Hz standard (BF_{10} = 0.26).

2.1.3.2. Mean normalised accuracy

Mean normalised accuracy for each comparison on the temporal (auditory, visual) and pitch generalisation tasks for the autistic and NT groups is displayed in Figure S4.

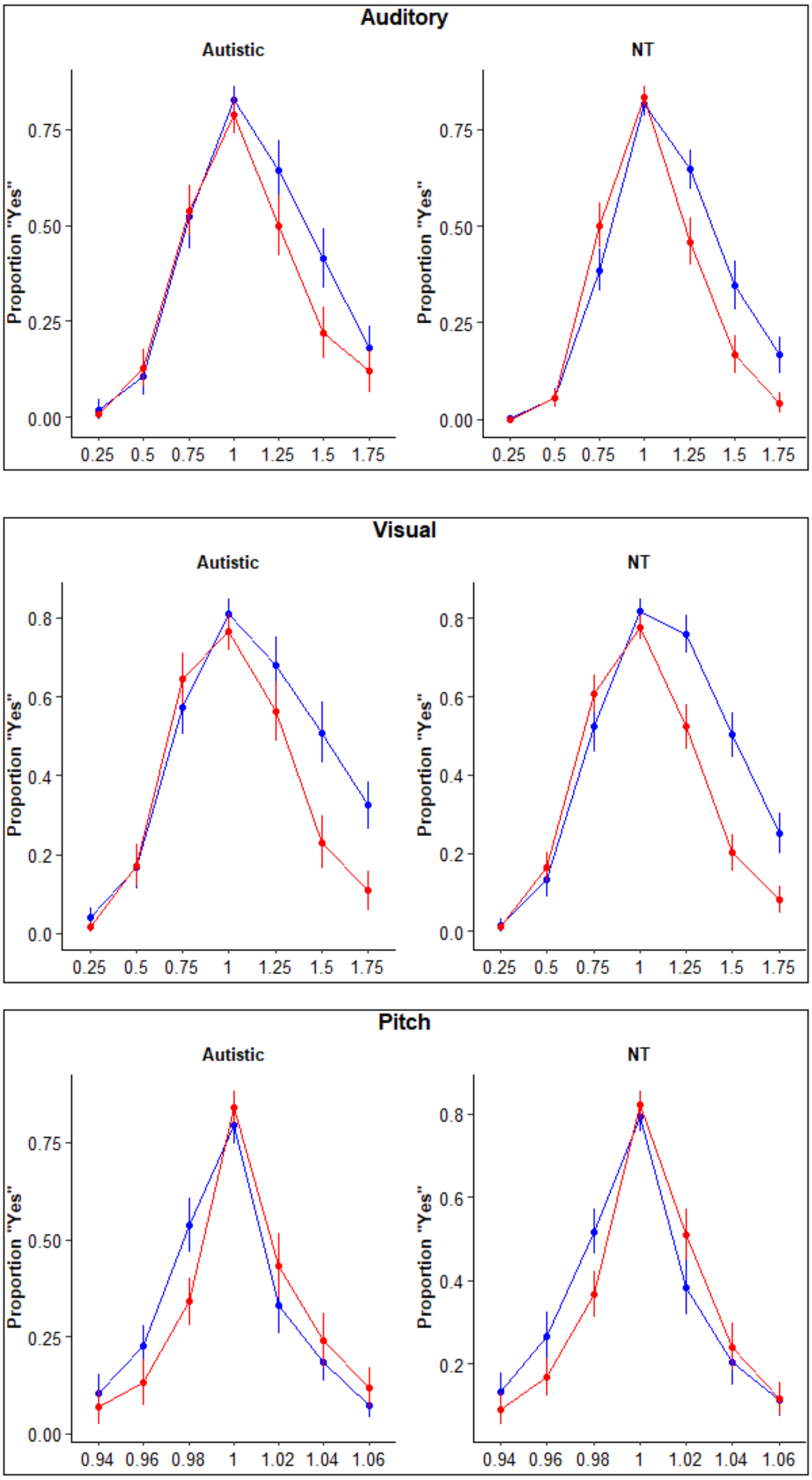

Figure S4. Mean proportion 'Yes' responses for each comparison in each condition of the Temporal Generalisation Tasks. 400ms (and 500hz in the pitch task) Standards are given in blue and 800ms (and 1000 Hz) are given in red. Error bars denote 95% CI.

As described in the pre-registration, the proportion of *yes* responses to each comparison condition was compared for each standard using a mixed ANOVA with Comparison (0.25, 0.50, 0.75, 1, 1.25, 1.50, 1.75) and Standard (400ms, 800ms) as within participant factors and Group (Autistic, NT) as the between participant factor. The autistic group tended to respond 'Yes' more frequently to comparisons in the auditory condition (a main effect of Group). There were no other differences between the groups.

In the auditory condition, there was a main effect of Comparison (F (6, 816) = 592, *p* < .001, η^2= .647) indicating that the mean proportion *yes* responses differed between the comparison durations. There was a main effect of Standard Duration (F (1, 136) = 64.67, *p* < .001, η^2= .017) reflecting that participants responded *yes* to comparisons more frequently in the 400ms standard condition (mean = .361 SD = .352) compared with the 800ms (mean = .307 SD = .344). There was a main effect of Group (F (1,136) = 7.94, *p* = .006, η^2 = .011) indicating that Autistic participants responded *yes* (mean = .362 SD = .351) more frequently than NT participants (mean = .318 SD = .347). There was a Comparison x Standard Duration interaction (F (6,816) = 20.89, *p* < .001, η^2 = .052). As can be seen in Figure S2 participants tended to overestimate the standard duration to a greater extent in the 400ms condition compared with the 800ms condition. There was no Group x Comparison interaction (F (6, 816) = 1.67, *p* = .125, η^2= .005) indicating that the shape of the temporal generalisation gradient was similar between the groups. There was no Group x Standard Duration interaction (F (1, 136) = 0.04, *p* = .838, η^2< .001) indicating the tendency to respond *yes* to comparison stimuli in the 400ms and 800ms standard duration conditions did not differ between the groups. There was no Group x Comparison x Standard Duration (F (6, 816) = 1.88, *p* = .082, η^2 = .005) indicating the relative shape of the temporal generalisation gradients between the standard conditions did not differ between the groups.

In the visual condition, there was a main effect of Comparison (F (6,780) = 479.57, *p* < .001, η^2 = .631) indicating that the proportion of *yes* responses differed between the comparison durations. There was a main effect of Standard Duration (F (1,130) = 176.78, *p* < .001, η^2 = .048) reflecting that participants responded *yes* more frequently to comparison stimuli in the 400ms standard condition (.433 SD = .349) compared with the 800ms (.341 SD = .338). There was no main effect of Group (F (1, 130) = 2.27, *p* = .135, η^2 = .003) reflecting that the tendency to respond *yes* did not differ between the groups. There was a Comparison x Standard Duration interaction (F (6, 780) = 41.98, *p* < .001, η^2 = .091) from Figure S2, it can be seen that participants tended to overestimate the standard duration to a greater extent in the 400ms condition than the 800ms condition. There was no Group x Comparison interaction (F (6, 780) = 1.03, *p* = .403, η^2 = .004) indicating that the shape of the temporal generalisation gradient was similar between the groups. There was no Group x Standard

Duration interaction (F (1, 130) = 1.26, *p* = 263, η^2< .001) indicating the tendency to respond *yes* to comparison stimuli in the 400ms and 800ms standard duration conditions did not differ between the groups. There was no Group x Standard Duration x Comparison (F (6, 780) = 1.63, *p* = .136, η^2 = .003) indicating the relative shape of the temporal generalisation gradients between the standard conditions did not differ between the groups.

In the pitch condition there was a main effect of Comparison (F (6, 780) = 479.57, *p* < .001, η^2= .632) indicating that the proportion of *yes* responses differed between the comparison frequencies. There was no main effect of Standard (F (1,121) = 1.86, *p* = .175, η^2 = .001) reflecting that the tendency to respond *yes* did not differ between the standard pitches. There was no main effect of Group (F (1, 121) = 1.69, *p* = .282, η^2= .271) reflecting that the tendency to respond *yes* did not differ between the groups. There was a Comparison x Standard interaction (F (6, 726) = 21.59¸ *p < .001,* η^2= .047) as can be seen in Figure S2, participants tended to underestimate the pitch of the standard in the 500hz condition. Group did not interact with Comparison (F (6, 726) = 1.25, *p* = .277, η^2 = .004) or Standard (F (1,121) = 0.09, *p* = .760, η^2 < .001) and there was no Group x Comparison x Standard interaction (F (6, 726) = 0.38, *p* = .891, η^2 = .001).

2.1.3.3. Coefficient of variation

Participants' coefficient of variation (mean – standard deviation) in response to each duration are represented in Figure S5

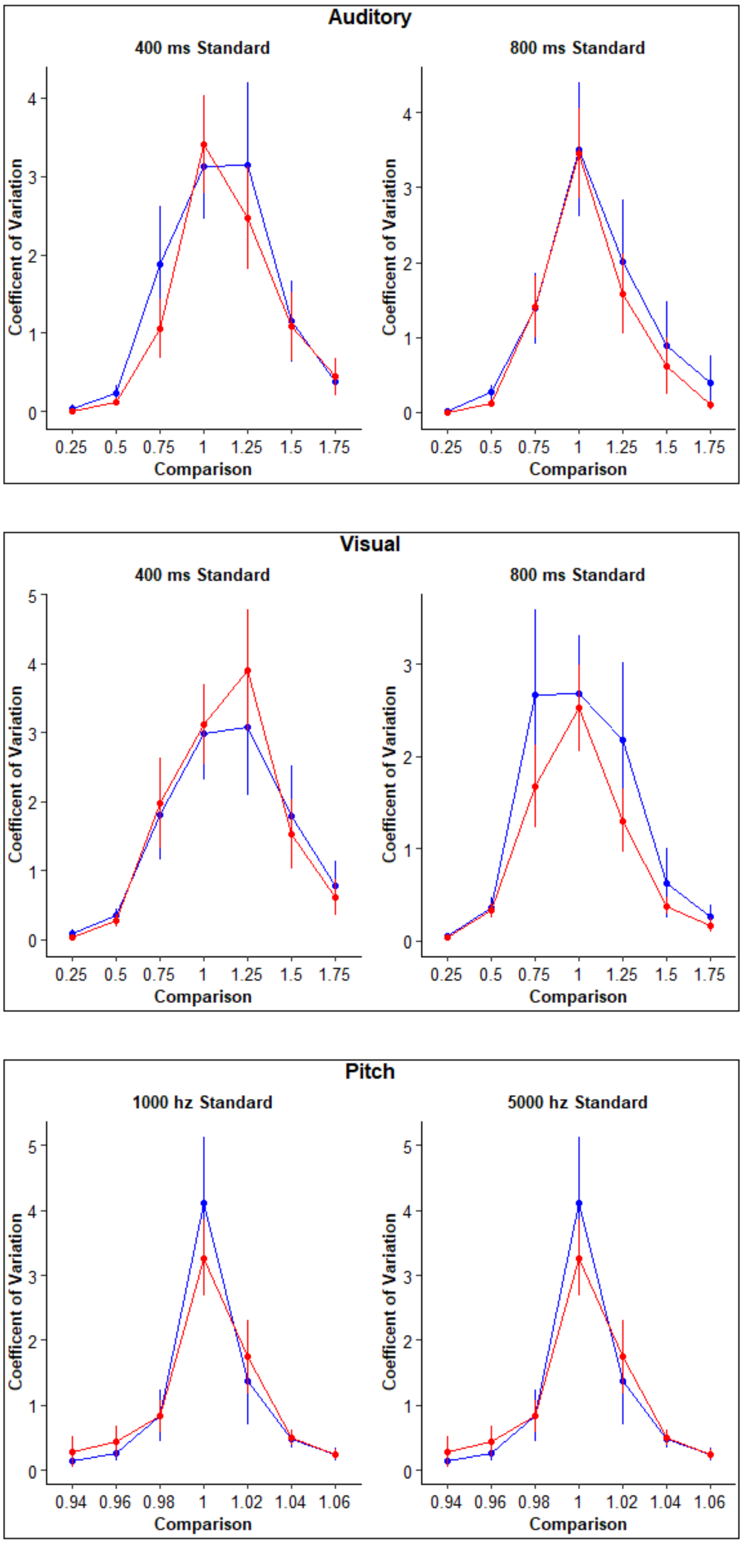

Figure S5. Mean coefficient of variation (mean – standard deviation) for each comparison for the Autistic (red) and NT (blue) groups. Error bars denote 95% CI.

2.1.4. Additional details of the Hierarchical Clustering on Principal Components

Correlation coefficients of each variable with the first three principal components (visually represented in Figure S3a) are given in Table S1.

Table S1. Correlation coefficient between each variable and the first three principal components

	PC1	PC2	PC3
Auditory TDT	.622	.334	.287
Visual TDT	.412	.552	.463
Auditory TOJ	.625	.277	.161
Visual TOJ	.488	.081	-.261
Auditory VE	.574	.162	-.601
Visual VE	.593	.308	-.557
Auditory TG	.658	.254	.204
Visual TG	.591	.187	.081
GSQ	.594	-.648	.149
DCD	.620	-.560	.032
AQ	.566	-.720	.077

There is further visualisation of the results of the HCPCA analysis in Figure S6.

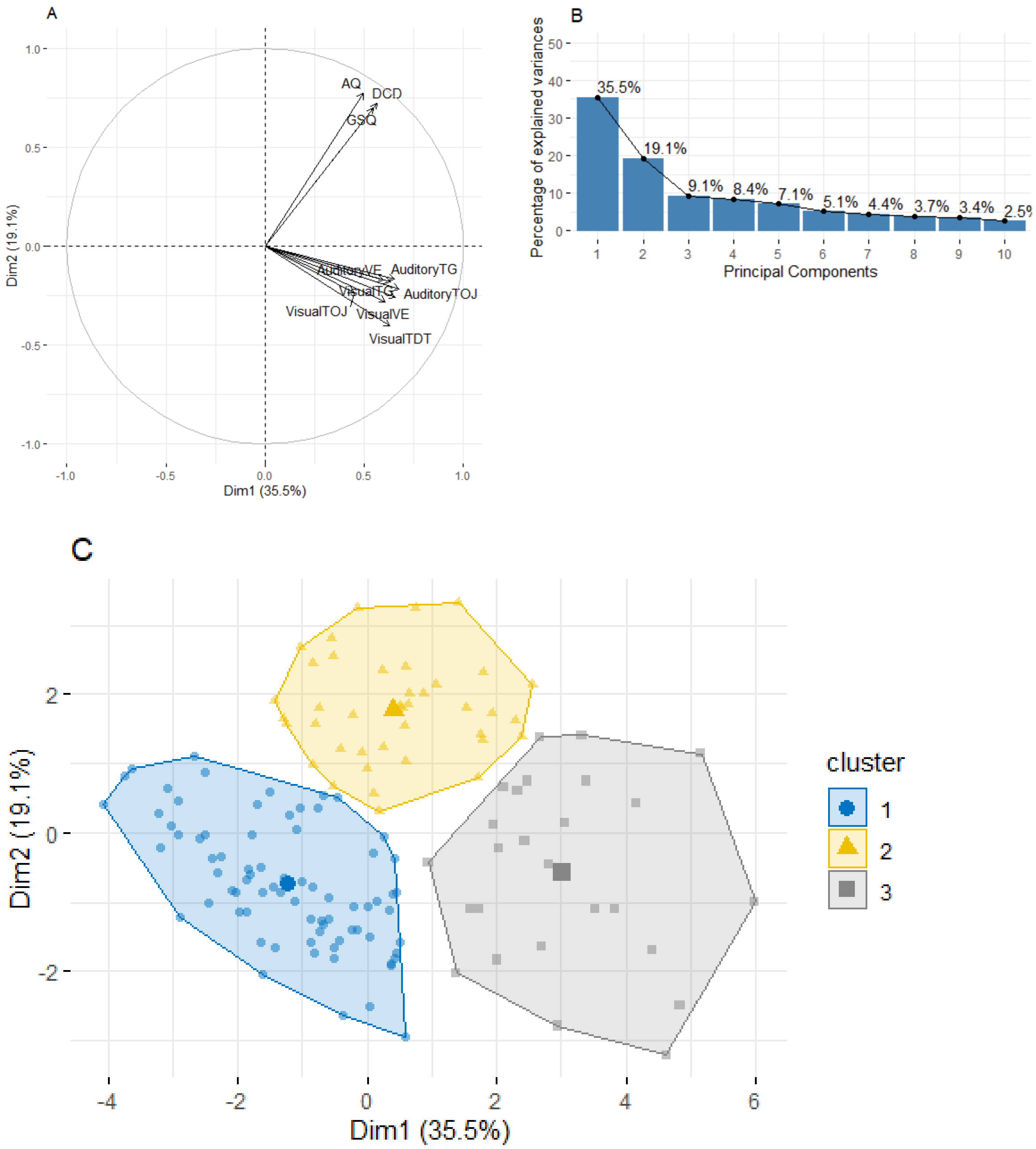

Figure S6 Further visualisations of the results of the hierarchical clustering on the principal components. A is a correlation circle representing the correlations between each variable and the first two principal components. B displays a Scree plot giving the ranked eigenvalues. The first three principal components explained 63.13% of the variance and the relative increase of further components is minimal. C displays the portioning of the clusters on the correlation circle. Data points represent individual participants, with the larger point representing the cluster mean.

Table S2 displays the v.test statistic which indicates where the mean of the cluster is statistically different from the group mean (α < .05; see Husson, Josse & Pages, 2010).

Table S2. ***Statistically significant v.test values for each cluster negative values denote where the cluster mean was lower than the group mean for that variable.***

	Cluster 1	**Cluster 2**	**Cluster 3**
	(2 Autistic, 74 NA)	(36 autistic, 4 NA)	(15 autistic, 11 NA)
Auditory TDT	v = -3.83 *p* < .001		v = 6.39, *p* < .001
Visual TDT	v = -3.32, *p* = .001	v = 2.18, *p* < .001	v = 6.86, *p* < .001
Auditory TOJ	v = -3.06, *p* = .002	v = -2.47, *p* = .014	v = 7.57, *p* < .001
Visual TOJ			v = 4.51, *p* < .001
Auditory VE	v = -3.65, *p* < .001		v = 5.02, *p* < .001
Visual VE	v = -2.45, *p* = .014		v = 4.65, *p* < .001
Auditory TG	v = -4.76, *p* < .001		v = 5.81, *p* < .001
Visual TG	v = -3.18, *p* = .001		v = 5.43, *p* < .001
GSQ	v = -8.55, *p* < .001	v = 6.67, *p* < .001	v = 3.27, *p* = .001
DCD	v = -9.07, *p* < .001	v = 7.60, *p* < .001	v = 2.86, *p* = .004
AQ	v = -9.58, *p* < .001	v = 8.11, *p* < .001	v = 2.87, *p* = .004

2.1.5. Repeating analysis with participants removed based on ADOS and AQ thresholds

2.1.5.1 ADOS

We did not exclude any participants based on ADOS or AQ scores. Of the ADOS scores we had access to (i.e. participants who were recruited from the community rather than directly from the NHS), the mean score was (9.67, SD = 2.93). There were three participants who score below the cut-off for autism (score of 7). We have repeated the analysis with these three participants (564, 5003, 539) removed and there are no meaningful changes in the findings of any of the tasks or the questionnaires.

2.1.5.2 AQ

There were four autistic participants (523,581,517,5009) who scored below the cut-off of 32 typically given on the AQ-50. Four participants in the NA group (781,785,721,103) scored over this cut-off. Again, we have repeated the entire analysis with these participants removed. Overall, there were no meaningful changes to the results apart from on the Verbal Estimation task where the main effect of group was just pushed below the α criterion (F (1, 129) = 4.29, p = .040, η^2= .024). The autistic group produced lower slopes overall than the NT group. The Bayes Factor for the between group comparisons of slopes in the auditory condition was $BF_{10} = 0.77$ and visual $BF_{10} = 0.84$ were not supportive of evidence in favour of between group differences.

3. Further discussion

An unexpected finding from the temporal generalisation task was that across the groups and modality conditions, sensitivity was increased in the 800 ms condition in comparison to the 400 ms. This invalidates scalar invariance which stipulates that variance (i.e. inverse sensitivity) scales linearly with increasing duration (Allan, 1998; Grondin, 2014; Haigh, Apthorp, & Bizo, 2020). Studies using the temporal generalisation task typically observe conformity to scalar invariance through superimposition of mean normalised accuracy for each comparison (see Derouet, Doyère, & Droit-Volet, 2019; Droit-Volet, 2002; Droit-Volet, Clément, & Wearden, 2001; Wearden, 1992; Wearden & Bray, 2001; Wearden, Denovan, Fakhri, & Haworth, 1997). However, a number of studies using temporal generalisation have observed data that does not conform to scalar invariance using the same analysis (Ferrara, Lejeune, & Wearden, 1997; Jones & Wearden, 2003, 2004; Lamotte, Droit-Volet, & Izaute, 2017), or when comparing Weber Fractions across durations (Grondin, 2010). In the present study, signal detection analysis also allowed the estimate of bias which suggested that participants were using a more liberal response criterion when responding to 400 ms standards in comparison to 800 ms (i.e. they had a greater tendency to respond 'yes'). In the typical temporal generalisation task, a single stimulus is presented on each trial and the participant makes a signal present or absent judgement (a Yes/No task), meaning a signal detection model may provide the best fit of the data (Stanislaw & Todorov, 1999) and provides a sensitive measure of performance taking observer bias into account (note also, that mean normalised accuracy data on the auditory and visual temporal generalisation tasks did not superimpose, particularly for durations longer than the standard. See supplementary materials S3). As scalar invariance is the hallmark of dedicated timing models such as SET there is value in using sensitive measures in larger samples to rigorously test the conditions under which scalar invariance holds on the temporal generalisation and other timing tasks.

4. References

Allan, L. G. (1998). The influence of the scalar timing model on human timing research. *Behavioural Processes*, *44*(2), 101–117. https://doi.org/10.1016/S0376-6357(98)00043-6

Derouet, J., Doyère, V., & Droit-Volet, S. (2019). The disruption of memory consolidation of duration introduces noise while lengthening the long-term memory representation of time in humans.

Frontiers in Psychology, *10*(APR), 1–10. https://doi.org/10.3389/fpsyg.2019.00745

Droit-Volet, Sylvie. (2002). Scalar timing in temporal generalization in children with short and long stimulus durations. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *55*(4), 1193–1209. https://doi.org/10.1080/02724980244000161

Droit-Volet, Sylvie, Clément, A., & Wearden, J. (2001). Temporal generalization in 3- to 8-year-old children. *Journal of Experimental Child Psychology*, *80*(3), 271–288. https://doi.org/10.1006/jecp.2001.2629

Ferrara, A., Lejeune, H., & Wearden, J. H. (1997). Changing Sensitivity to Duration in Human Scalar Timing: An Experiment, a Review, and Some Possible Explanations. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *50*(3), 217–237. https://doi.org/10.1080/713932654

Grondin, S. (2010). Unequal Weber fractions for the categorization of brief temporal intervals. *Attention,Perception & Psychophysics*, *72*(5), 1422–1430. https://doi.org/10.3758/APP

Grondin, Simon. (2014). About the (non)scalar property for time perception. In H. Merchant & V. de Laufente (Eds.), *Neurobiology of Interval Timing* (pp. 17–29). New York, USA: Springer.

Haigh, A., Apthorp, D., & Bizo, L. A. (2020). The role of Weber's law in human time perception. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-020-02128-6

Jones, L. A., & Wearden, J. H. (2003). More is not necessarily better: Examining the nature of the temporal reference memory component in timing. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *56*(4), 321–343. https://doi.org/10.1080/02724990244000287

Jones, L. A., & Wearden, J. H. (2004). Double standards: Memory loading in temporal reference memory. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *57*(1), 55–77. https://doi.org/10.1080/02724990344000088

Lamotte, M., Droit-Volet, S., & Izaute, M. (2017). Confidence judgment in a temporal generalization task: Accuracy and sensitivity to task difficulty. *Annee Psychologique*, *117*(3), 275–298. https://doi.org/10.4074/S0003503317003025

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *3*(I), 37–149.

Wearden, J. H., Denovan, L., Fakhri, M., & Haworth, R. (1997). Scalar timing in temporal generalization in humans with longer stimulus durations. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*(4), 502–511. https://doi.org/10.1037/0097-7403.23.4.502

Wearden, J. H. (1992). Temporal generalization in human. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*(2), 134–144.

Wearden, J. H., & Bray, S. (2001). Scalar timing without reference memory? Episodic temporal generalization and bisection in humans. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *54*(4), 289–309.